Simple Method for Locating Possible Ligand Binding Sites on Protein Surfaces

ANDREY A. BLIZNYUK,* JILL E. GREADY

Computational Molecular Biology and Drug Design Group, John Curtin School of Medical Research, The Australian National University, Canberra, Australia

Received 4 September 1998; accepted 22 February 1999

ABSTRACT: A new, fast, and easy-to-implement method, van der Waals–fast Fourier transform (vdW-FFT), for locating possible binding sites on the surface of a protein was developed and tested on a set of 15 different enzyme–ligand complexes. The method scans the whole protein surface and possible ligand orientations in order to find the best geometrical match, which corresponds to the minimum of the modified vdW energy. Two different grids, fine and coarse, and two sets of MM parameters, from the OPLS and AMBER-94 force fields, were used. The method has been shown to work accurately on the fine grid. On the coarse grid, the vdW-FFT method failed only on two complexes. The C program implementing the method and test set of proteins is available free on our web site: http://biocomp.anu.edu.au/~aab. © 1999 John Wiley & Sons, Inc. J Comput Chem 20: 983–988, 1999

Keywords: molecular docking; protein binding sites; fast Fourier transform (FFT); van der Waals (vdW) interactions; enzyme active sites

Introduction

Understanding the mechanisms governing association of proteins with their ligands or other proteins (molecular docking) is one of the most important topics in molecular biology. A general solution to the problem of determining the binding modes of biomolecules from knowledge of their molecular structure will provide the necessary foundation for a vast number of potential practical applications. These range from drug design and vaccine development to the engineering of novel proteins with predefined functionality.

The molecular docking of ligands to proteins requires two different issues to be addressed. First, the positions where the ligand will bind to the protein have to be determined. Second, the energies of binding and actual orientation of the ligand need to be estimated. In many cases, the locations

* *Present address:* A. A. Bliznyuk, ANU Supercomputer Facility, Information Technology Services, The Australian National University, Canberra, ACT 0200, Australia; e-mail: Andrey.Bliznyuk@anu.edu.au

Correspondence to: J. E. Gready; e-mail: jill.gready@anu.edu.au

Contract/grant sponsor: ANU (Strategic Development Grant)

CCC 0192-8651 / 99 / 090983-06

of the binding sites are known exactly from crystallographic or NMR data or approximately from mutagenesis experiments. As a result, most efforts concentrate on the second task.[1–3] However, as more protein structures become available, the general approach to the location of possible binding sites becomes increasingly important.[1]

A number of methods have been proposed to deal with the problem of locating possible binding sites of a ligand.[4–11] These approaches can be divided into two groups.

The first group includes such methods as GRID[4] and the so-called ''pocket-finder'' method.[6] These methods use some kind of energy estimation to locate possible binding sites. Thus, the ability of these methods to find a binding site depends on the accuracy of the function used for the energy estimation. The authors of the GRID program[4] used a function similar to the molecular mechanics energy function. Only one successful application of this program to this general docking problem has been published.[5] The plot of energy difference between binding and nonbinding tautomers of a ligand was used for locating the binding site. It is doubtful that this approach can be used for the general task of selecting possible binding sites. A completely different set of empirical functions was chosen and parameterized by the authors of ref. 6. The parameters were optimized to reproduce known experimental binding constants,[6] which are available for active sites only. The ''pocket-finder'' method[6] was shown to work quite well on a set of 10 enzyme–ligand complexes. However, as the authors pointed out, it is not clear whether the method can be applied successfully to identifying binding sites that do not resemble active sites, such as on the exterior of a protein.

The second group includes variants of a pure geometrical approach to the selection of possible binding sites. Several approaches based on the idea of geometric complementarity between ligand and an enzyme binding site have been proposed.[7–12] All these methods appear to be able to select several dozen possible binding sites, usually including the actual binding site. However, it should be noted that the methods[7–12] were developed for docking of protein–protein complexes and have not been tested extensively for complexes of proteins with small ligands. For example, only three such complexes were included in the test set in ref. 8 and only two in refs. 10 and 11. Identifying possible docking sites of small ligands is a quite different problem from searching for possible aggregation sites between two large enzyme molecules. For example, our attempt to use the method described in ref. 11 for finding possible binding sites of small ligands on the surface of a protein failed.[13]

In this article, we describe a new and very simple approach, vdW-FFT (van der Waals–fast Fourier transform), to identify possible binding sites. This method is intermediate between the two groups just described. It uses a modified vdW energy as the criterion for a good geometrical match.

Method

The vdW energy using OPLS[14] or AMBER-94[15] force fields is computed using a 6–12 potential:

$$E_{\mathrm{vdW}} = \sum_{i>j} A_{ij}/R_{ij}^{12} - B_{ij}/R_{ij}^{6} \qquad (1)$$

where R_{ij} is the interatomic distance and the summation is done over all interacting pairs of atoms. Parameters A_{ij} and B_{ij} are determined as follows:

$$\begin{aligned} A_{ij}(\mathrm{Amber}) &= \epsilon_{ij}\sigma_{ij}^{12} & A_{ij}(\mathrm{OPLS}) &= 4\epsilon_{ij}\sigma_{ij}^{12} \\ B_{ij}(\mathrm{Amber}) &= 2\epsilon_{ij}\sigma_{ij}^{6} & B_{ij}(\mathrm{OPLS}) &= 4\epsilon_{ij}\sigma_{ij}^{6} \\ \epsilon_{ij}(\mathrm{Amber}) &= (\epsilon_i\epsilon_j)^{1/2} & \epsilon_{ij}(\mathrm{OPLS}) &= (\epsilon_i\epsilon_j)^{1/2} \\ \sigma_{ij}(\mathrm{Amber}) &= (\sigma_i + \sigma_j) & \sigma_{ij}(\mathrm{OPLS}) &= (\sigma_i\sigma_j)^{1/2} \end{aligned} \qquad (2)$$

In eqs. (2), ϵ_i and σ_i are force field parameters. A good geometrical match occurs when the vdW energy is negative. Indeed, when the distance between ligand and protein atoms is short, the vdW energy is positive. As the distance increases, it becomes negative and goes to zero as the distance increases further. The better the interaction match between protein and ligand, the lower the vdW energy. A search for a minimum on the energy hypersurface is, in general, not a trivial task. However, if the energy can be expressed in the form of eq. (3), fast Fourier transform (FFT) can be used to search for a minimum and this simplifies the task considerably[11,16]:

$$E(p,q,t) = \sum_k \sum_l \sum_m V(k,l,m) * W(k+p, l+q, m+t) \qquad (3)$$

where V and W are some values, defined on a three-dimensional grid, and E is the energy. We used two separate three-dimensional grids of size $N \times N \times N$ to compute the vdW energy between protein and ligand. Points on one grid are assigned to a modified vdW repulsion energy:

$$V(k, l, m) = \sum_i \sqrt{A_{ii}} / (R + d/2)^{12} \qquad (4)$$

where (k, l, m) is a grid-point index, R is the distance between the grid point and an atom i, and d is the grid spacing. Summation is over all atoms within the cutoff distance from point (k, l, m). We used 8 Å for the cutoff distance. Beyond 8 Å, the vdW energy becomes small enough to be ignored. To avoid large values of $V(k, l, m)$, which may occur if $(R + d/2)$ is small, we use 1 Å whenever $(R + d/2)$ is less than 1 Å. The A_{ii} are computed as in eqs. (5) and (6) for OPLS and AMBER-94 force fields, respectively:

$$A_{ii}(\text{OPLS}) = 4\epsilon_i \sigma_i^{12} \qquad (5)$$

$$A_{ii}(\text{Amber}) = 2048\,\epsilon_i \sigma_i^{12} \qquad (6)$$

Points on the other grid were assigned to the vdW attraction energy, defined as:

$$W(k, l, m) = -\sum_i \sqrt{B_{ii}} / R^6 \qquad (7)$$

Values of B_{ii} are computed using eqs. (8) and (9) for OPLS and AMBER-94 force fields, respectively:

$$B_{ii}(\text{OPLS}) = 4\epsilon_i \sigma_i^{6} \qquad (8)$$

$$B_{ii}(\text{Amber}) = 128\,\epsilon_i \sigma_i^{6} \qquad (9)$$

The ligand's vdW parameters were projected onto the nearest eight points of grids with the same dimensions as the enzyme grids using:

$$V_L(k, l, m) = \sum_i (1 - a/d)(1 - b/d) \times (1 - c/d)\sqrt{A_{ii}} \qquad (10)$$

$$W_L(k, l, m) = \sum_i (1 - a/d)(1 - b/d) \times (1 - c/d)\sqrt{B_{ii}} \qquad (11)$$

where the values of a, b, and c are distances along the x-, y-, and z-coordinates from atom i to the grid point (k, l, m). We found that this simple linear projection worked well and did not experiment with others.

The interaction energy of the enzyme with the ligand molecule shifted over the distance (p^*d, q^*d, t^*d) is a simple multiplication and summation over grid points:

$$E(p, q, t) = \sum_k^N \sum_l^N \sum_m^N \{V(k, l, m) * V_L(k + p, l + q, m + t) + W(k, l, m) \times * W_L(k + p, l + q, m + t)\} \qquad (12)$$

Eq. (12) is similar to eq. (3) and, consequently, the FFT search for the minimum can be used for this equation. The only difference is that it requires two FFT operations, as described elsewhere.[16]

The reason for the use of parameter $d/2$ in eq. (4) is clear from eq. (12). It makes the repulsive grid smoother, thus reducing errors due to the discrete nature of the grid. We find that the use of $d/2$ in eq. (4) is necessary to produce meaningful results with small ligands and a moderate grid size (N = 64).

The use of eqs. (6) and (9) for the AMBER-94 force field introduces additional error to the computed vdW energy compared with the eq. (2). However, as was shown in ref. 17, this error is small and can be neglected.

A similar type of vdW energy projection onto grids has been used in ref. 16 for the purpose of calculating the exact MM energies. To obtain an accurate estimation of the interaction energies, a very fine grid was utilized and parameter $d/2$ was not used in eq. (4). The goal of our approach is to select possible binding sites, using the idea of geometrical complementarity between the receptor site and the ligand. Hence, accurate evaluation of the vdW energy is not needed. On the other hand, the use of a simple geometrical projection, as was proposed in ref. 11, did not allow us to select relevant candidate sites. Therefore, we decided to employ the modified form of the vdW energy, as described in eqs. (4)–(12).

In order to scan all possible orientations of the ligand with respect to the enzyme, we vary three Euler angles that define the orientation, by fixed steps of 20°, similarly to the method in ref. 11. This results in the generation of 360/20 * 360/20 * 180/20 = 2916 possible rotations, only 2628 of which correspond to dissimilar orientations. Hence, the procedures of projecting the ligand onto the grids and the FFT-based search of minima have to be repeated 2628 times. Fortunately, the FFT procedure is fast and the whole operation requires approximately 40 minutes on the SGI Power Indigo^2 (R8000) workstation for grids with the number of grid points (N) equal to 64.

Results

To obtain a representative test set we used protein–ligand complexes from sets employed previously for similar purposes.[9,18,19] Data for the 15 chosen enzyme–ligand complexes are summarized in Table I.

Coordinates of all the enzyme–ligand complexes were obtained from the Protein data bank.[20] The missing hydrogens were added using the INSIGHTII program from MSI.[21] The geometries of the enzymes and ligands were fixed during the computations. Atomic charges on ligands for the MM computations were generated by fitting the electrostatic potential obtained with the 6-31G* basis set using the Merz–Kollman procedure[22,23] as implemented in the GAUSSIAN94 program.[24] Atomic charges for protein atoms were obtained from the standard OPLS[14] or AMBER-94[15] parameters. These charges were used during the rigid-body geometry MM energy optimization as described in what follows.

To test the stability of the method we performed calculations with two different grids: a fine grid with N = 120 and coarse grids with N = 64 and 70. The grid step d was approximately 0.5 and 1 Å for the fine and coarse grids, respectively. Also, two different sets of vdW parameters, from the OPLS[14] and AMBER-94[15] force fields, were used. Calculations with the grid sizes of 120 and 70 were done on the Fujitsu VPP300 supercomputer, the computation time for each enzyme being approximately 60 and 20 minutes, respectively. The calculations with the grid size of 64 were performed on the SGI Indigo^2 R8000 workstation, the time being approximately 40 minutes for each enzyme. The ranking of the ligand position in the active site is given in Table II. We consider the ligand to be in the correct position if the root-mean-square deviation (RMSD) of its coordinates is < 2 Å from their crystallographic positions. In most cases the RMSD is < 1 Å. In all cases where the RMSD was larger than 1 Å, but smaller than 2 Å, a simple MM rigid-body optimization reduced the RMSD to below 1 Å.

In all the test systems, the method was able to locate the right binding position of a ligand and rank it in the top 15 when the fine grid was used. In most cases, the correct position received rank 1. Computations with the coarse grids (N = 64 or 70) failed for the 1GST complex (OPLS parameters) and for the 1LDM complex (AMBER-94 parameters). Also, the right position of the ligand in the 6RSA complex was ranked 70 when the OPLS parameters were used. Overall, the vdW-FFT method works very well with fine grids, but becomes less reliable when the coarse grids are used.

Discussion

The vdW-FFT method we have introduced scans the whole protein surface and possible ligand orientations in order to find the best geometrical match, which corresponds to the minimum of the

TABLE I.
Enzyme – Ligand Complexes Used in Calculations.

Number	PDB File Name	Enzyme	Ligand
1	1GST	Glutathione-*S*-transferase	Glutathione
2	1LDM	Lactate dehydrogenase	NAD
3	2CTC	Carboxypeptidase A	*L*-phenyl-lactate
4	1STP	Streptavidin	Biotin
5	2GBP	D-Galactose / D-glucose-binding protein	β-D-glucose
6	2IGF	IgG1 Fab fragment	Myohemerythrin
7	2PHH	*p*-Hydroxybenzoate hydrolase	Adenosine-5-diphosphoribose
8	3CPA	Carboxypeptidase A	Glycyl-L-tyrosine
9	3DFR	Dihydrofolate reductase	NADPH
10	3DFR	Dihydrofolate reductase	Methotrexate
11	3PTB	β-Trypsin	Benzamidine
12	4DFR	Dihydrofolate reductase	Methotrexate
13	4MBN	Metmyoglobin	Heme
14	6RSA	Ribonuclease A	Uridine
15	4PHV	HIV-1 protease	Inhibitor

TABLE II.
Rank from vdW – FFT Calculations of Correct Positioning of Ligand in Binding Site of Structures Given in Table I.

		OPLS Parameters		AMBER94 Parameters	
Number	PDB File	Grid 120 × 120 × 120	Grid 64 × 64 × 64	Grid 120 × 120 × 120	Grid 70 × 70 × 70
1	1GST	3	> 100	1	1
2	1LDM	1	1	1	> 100
3	2CTC	4	16	2	9
4	1STP	1	2	1	2
5	2GBP	1	1	1	4
6	2IGF	1	1	1	1
7	2PHH	1	1	1	1
8	3CPA	1	4	5	4
9	3DFR	1	1	1	1
10	3DFR	1	1	1	1
11	3PTB	5	2	8	10
12	4DFR	1	1	1	1
13	4MBN	1	1	1	1
14	6RSA	15	70	2	3
15	4PHV	1	1	1	1

modified vdW energy. The aim of the method is to reduce a very large number of possible binding sites to a very few, which can then be studied using more sophisticated energy functions. For example, in the latter regard, we successfully used a combination of molecular mechanics energy with solvation energy obtained either by the DELPHI program[25] or by using our PRFM method[26] to study pterin docking to dihydrofolate reductases,[27] starting from vdW-FFT guesses. It should be noted, however, that the application of such a combination might lead to incorrect results. For example, our attempt to improve the rank order of the possible docking sites obtained for the test enzymes studied in the current article failed. It seems that more accurate energy estimation based on the free energy evaluation[28] is necessary in this case.

The vdW-FFT method uses projections of the modified vdW energy onto grids. Thus, the accuracy of the method should depend on the grid size employed. When the grid is coarse, the method fails. The results presented here show two such failures. First, the method ranked the complex of glutathione–*S*-transferase with glutathione (1GST) below 100 when OPLS parameters were used. Glutathione is a small molecule and consists of only 26 atoms, but it binds in a very large binding pocket and approximately half of it is exposed to the solvent. When the coarse grid is used, the projection of the ligand loses some details that are necessary to describe binding correctly. The correct ranking was obtained with the AMBER-94 parameters, because the aliphatic hydrogen atoms present in AMBER-94 force field made the projection slightly more accurate.

Second, the method failed for the 1LDM complex with the AMBER-94 parameters. Lactate dehydrogenase monomer is large enzyme of more than 5000 atoms. As a result, the grid step for the coarse grid was 1.3 Å and the binding pocket for the NAD cofactor molecule was not rendered accurately. NAD binds in a very tight pocket and the aliphatic hydrogens present in the AMBER-94 parameters contributed additional repulsion making the energy of the correct binding mode too high. This did not happen when the OPLS parameters were used.

The vdW-FFT method uses only the vdW energy to rank possible binding sites. It would not be difficult to include Coulomb interactions that can be obtained either as a projection of the electrostatic potential onto the grid or from solution of the Poisson–Boltzmann equation using such programs as DELPHI.[25] However, our preliminary investigations indicate this did not improve results, especially for the coarse grids. Also, this would require an additional FFT procedure, which increases the computation time by 30% compared with the current version.

In its current form, the vdW-FFT method is not applicable directly to flexible ligands. However, the discrete nature of the grid and smoothed vdW

projection given by eq. (4) can accommodate possible small geometry changes of both the enzyme and the ligand during binding. If the ligand has several distinct conformations, then the vdW-FFT method can be applied for each separate conformation. Another way of dealing with flexible ligands would be to separate them into several rigid parts and apply the vdW-FFT method to each part. Fortunately, the vdW-FFT is fast enough to allow such computations.

Acknowledgments

Computer time grants from the Australian National University Supercomputer Facility are gratefully acknowledged.

References

1. Balbes, L. M.; Mascarella, S. W.; Boyd, D. B. In: Lipkowitz, K. B.; Boyd, D. B., eds. Reviews in Computational Chemistry, Vol. 5; VCH: New York, 1994, p 337.
2. Böhm, H. J.; Klebe, G. Angew Chem Int Ed Engl 1996, 35, 2589.
3. Murcko, M. A. In: Lipkowitz, K. B.; Boyd, D. B., eds. Reviews in Computational Chemistry, Vol. 11; Wiley-VCH: New York, 1997, p 1.
4. Bobbyer, D. N. A.; Goodford, P. J.; McWhinnie, P. M.; Wade, R. C. J Med Chem 1989, 32, 1083.
5. Reynolds, C. A.; Wade, R. C.; Goodford, P. J. J Mol Graph 1989, 7, 103.
6. Ruppert, J.; Welch, W.; Jain, A. N. Prot Sci 1997, 6, 524.
7. Kunz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Venkataraghavan, R. J Med Chem 1986, 29, 2149.
8. Lin, S. L.; Nussinov, R. J. J Mol Graph 1996, 14, 78.
9. Fischer, D.; Lin, S. L.; Wolfson, H. L.; Nussinov, R. J. J Mol Biol 1995, 248, 459.
10. Kasinos, N.; Lilley, G. A.; Subbarao, N.; Haneel, I. Prot Eng 1992, 5, 69.
11. Katchalski-Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A. A.; Aflalo, C.; Vakser, I. A. Proc Natl Acad Sci USA 1992, 89, 2195.
12. Vakser, I. A.; Aflalo, C. Proteins 1994, 20, 320.
13. Bliznyuk, A. A.; Gready, J. E. Unpublished results.
14. Jorgensen, W. L.; Tirado-Rives, J. J Am Chem Soc 1998, 110, 1664.
15. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Cadwell, J. W.; Kollman, P. A. J Am Chem Soc 1995, 117, 5179.
16. Harrison, R. W.; Kourinov, I. V.; Andrews, L. C. Prot Eng 1994, 7, 359.
17. Luty, B. A.; Wasserman, Z. R.; Stouten, P. F. W.; Hodge, C. N.; Zacharias, M.; McCammon, J. A. J Comput Chem 1995, 16, 454.
18. Rarey, M.; Wefing, S.; Lengaur, T. J Comput-Aid Mol Des 1996, 10, 41.
19. Gschwend, D. A.; Kuntz, I. D. J Comput-Aid Mol Des 1996, 10, 123.
20. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. D.; Kennard, O.; Shimanouchi, T.; Tasumi, M. J Mol Biol 1977, 112, 535.
21. INSIGHT II, Version 95, October 1995. Molecular Simulations Inc., 9685 Scranton Road, San Diego, CA 92121-3752.
22. Singh, U. C.; Kollman, P. A. J Comput Chem 1984, 5, 129.
23. Besler, B. H.; Merz, K. M., Jr.; Kollman, P. A. J Comput Chem 1990, 11, 431.
24. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Gill, P. M. W.; Johnson, B. G.; Robb, M. A.; Cheeseman, J. R.; Keith, T. A.; Peterson, G. A.; Montgomery, J. A.; Raghavachari, K.; Al-Laham, M. A.; Zakrzewski, V. G.; Ortiz, J. V.; Foresman, J. B.; Cioslowski, J.; Stefanov, B. B.; Nanayakkara, A.; Challacombe, M.; Peng, C. Y.; Ayala, P. Y.; Chen, W.; Wong, M. W.; Andres, J. L.; Replogle, E. S.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Binkley, J. S.; Defrees, D. J.; Baker, J.; Stewart, J. P.; Head-Gordon, M.; Gonzalez, C.; Pople, J. A. GAUSSIAN-94 (Revision C.2); Gaussian Inc.: Carnegie, PA.
25. DELPHI Program, Biosym Technologies, Version 2.5, 1993. Molecular Simulations Inc.: San Diego, CA.
26. Bliznyuk, A. A.; Gready, J. E. J Phys Chem 1995, 99, 14506.
27. Bliznyuk, A. A.; Gready, J. E. J Comput-Aid Mol Des 1998, 12, 325.
28. Kollman, P. A. Curr Opin Struct Biol 1994, 4, 240.